In this part of the course we will learn how to solve systems of linear equations of the form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots & \vdots \\ \vdots \quad\quad \ddots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

where a_{ij} and b_i are real numbers, and x_j are the unknowns, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$. We will write this system in the matrix-vector form

$$Ax = b, \qquad (*)$$

where A is the $n \times n$ matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

and b is an $n-$dimensional vector,

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbf{R}^n$$

Finding solution of (*) means that we find the result of multiplying of the inverse matrix A^{-1} with b:

$$x = A^{-1}b.$$

The inverse A^{-1} to a matrix A is, by definition, any matrix B which satisfies $AB = BA = I$, where I is the identity matrix

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & & & \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

If a matrix A has an inverse, it is called *nonsingular*.

Theorem. The following statements are equivalent:

1) The matrix A is nonsingular;
2) $\det A \neq 0$;
3) For every vector b the system $Ax = b$ has at least one solution;
4) For every vector b the system $Ax = b$ has exactly one solution.
5) $\operatorname{rank} A = \operatorname{rank} [A\, b] = n$.

Example: Consider the 2×2 system

$$\begin{aligned} ax + by &= e \\ cx + dy &= f \end{aligned}$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} e \\ f \end{pmatrix}.$$

We have

$$\det A = ad - bc \quad \text{and } A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Then

$$x = \frac{de - bf}{ad - bc} \quad y = \frac{-ce + af}{ad - bc}.$$

$\operatorname{rank} A =$ the number of linearly independent columns (rows).
Linear independence of vectors:

$$\sum_{i=1}^{r} \alpha_i x_i \quad \Rightarrow \quad \sum_{i=1}^{r} \alpha_i^2 = 0$$

Example:

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

$$\operatorname{rank} A = \operatorname{rank} [A\, b] = 1$$

Infinitely many solutions.

Theorem. If $\operatorname{rank} A = \operatorname{rank} [A\, b]$ then there is at least one solution of $Ax = b$. Specifically, if $\operatorname{rank} A = \operatorname{rank} [A\, b] = n$ there is exactly one solution, and if $\operatorname{rank} A = \operatorname{rank} [A\, b] = r < n$, then the set of solution is an affine manifold of dimension $n - r$.

1. Gaussian Elimination

It is easier to solve a system of equations $Ax = b$ if the matrix A has a special structure:

- The matrix A is diagonal: solve n independent equations of one variable .

  Example:

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \Rightarrow \quad x_1 = 2, \quad x_2 = 0.5 \ .$$

- The matrix A is upper triangular: backward substitution.

  Example:

$$\begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \Rightarrow \quad x_2 = .5, \quad x_1 = 2 - 3 \cdot 0.5 = 0.5 \ .$$

- The matrix A is lower triangular: forward substitution.

  Example:

$$\begin{pmatrix} 1 & 0 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \Rightarrow \quad x_1 = 2, \quad x_2 = (1 - 3 \cdot 2)/2 = -2.5 \ .$$

Gaussian elimination: use elementary row transformations to reduce $Ax = b$ to an equivalent upper triangular system and then apply backward substitution.

Elementary row operations:

- multiply an equation by a nonzero number;
- add (subtract) two equations;
- exchange two equations.

Performing elementary row operation on a linear system, transform the system in an equivalent system in the sense that it has the same solution.

In Gaussian elimination we use the first two properties.

Example:

$$\begin{array}{rcrcl} x_1 & + & 2x_2 & = & 0 \\ 2x_1 & + & 2x_2 & = & 3 \end{array} \qquad (E1)$$

The goal is to make zero the coefficient (now 2) in from of x_1 in the second equation. We can do this by multiplying the first equation by a number such that the coefficients in front of x_1 in the first and in the second equations are the same, and then subtract the first equation from the second, putting the result as a second equation.

Multiplying the first equation by 2 we obtain

$$\begin{array}{rcrcl} 2x_1 & + & 4x_2 & = & 0 \\ 2x_1 & + & 2x_2 & = & 3 \end{array} \qquad (E2)$$

Now we write the first equation in (E1) (without changes) and in place of the second equation we write the equation obtained by subtracting the first equation from the second in (E2):

$$\begin{array}{rcrcl} x_1 & + & 2x_2 & = & 0 \\ & & -2x_2 & = & 3 \end{array}$$

The obtained system is in the upper triangular form; hence, we can solve the second equation obtaining $x_2 = -3/2$. Then, substituting in the first equation we get $x_1 = -2x_2 = 3$.

Example:

$$\begin{array}{rcrcrcll} x_1 & + & 2x_2 & + & x_3 & = & 0 & (E1) \\ 2x_1 & + & 2x_2 & + & 3x_3 & = & 3 & (E2) \\ -x_1 & - & 3x_2 & & & = & 2 & (E3) \end{array}$$

Step 1: To eliminate x_1 from (E2) we subtract 2 times (E1) from (E2). To eliminate x_1 from (E3) we subtract -1 times (E1) from (E3).

$$\begin{array}{rcrcrcll} x_1 & + & 2x_2 & + & x_3 & = & 0 & (E1) \\ & - & 2x_2 & + & x_3 & = & 3 & (E2) \\ & - & x_2 & + & x_3 & = & 2 & (E3) \end{array}$$

Step 2: To eliminate x_2 from (E3) we subtract 1/2 times (E2) from (E3).

$$\begin{array}{rcrcrcll} x_1 & + & 2x_2 & + & x_3 & = & 0 & (E1) \\ & - & 2x_2 & + & x_3 & = & 3 & (E2) \\ & & & & \frac{1}{2}x_3 & = & \frac{1}{2} & (E3) \end{array}$$

Step 3: Backward substitution. Solve first for x_3, then for x_2 and finally for x_1:

$$\begin{aligned} x_3 &= 1 \\ -2x_2 + 1 &= 3 \\ x_2 &= -1 \\ x_1 + 2(-1) + 1 &= 0 \\ x_1 &= 1 \end{aligned}$$

Gaussian elimination: the general case

The system $Ax = b$ is represented in the *augmented* matrix form

$$(A|b) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & & \ddots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array} \right)$$

At the k-th iteration we have the matrix

$$(A^{(k)}|b^{(k)}) = \left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ & a_{22}^{(2)} & & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ & & \ddots & & & \vdots \\ & & & a_{kk}^{(k)} \quad \cdots & a_{kn}^{(k)} & b_k^{(k)} \\ & & & \vdots \quad \ddots & & \vdots \\ & & & a_{n1}^{(k)} \quad \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{array} \right)$$

The k-th step consists in transforming the matrix such that $a_{j,k}^{(k+1)}$ become zero for $j = k+1, \cdots, n$. Suppose that

$$a_{kk}^{(k)} \neq 0\,. \tag{P}$$

Then transform the matrix above in the following way:

1^0. For $j = k+1, k+2, \cdots, n$: Compute the *multipliers*:

$$m_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}} \quad j = k+1, k+2, \cdots, n$$

2^0. For $i = k+1, k+2, \cdots, n$: Determine the new components:

$$a_{ji}^{(k+1)} = a_{ji}^{(k)} - m_{jk} \cdot a_{kj}^{(k)}$$

$$b_j^{(k+1)} = b_j^{(k)} - m_{jk} \cdot b_k^{(k)}$$

End of 2^0. End of 1^0.

When the $n-1$st step is completed, we obtain a system in the upper triangular form

$$(A^{(n)}|b^{(n)}) = \left(\begin{array}{cccc|c} a_{11}^{(n)} & a_{12}^{(n)} & \cdots & a_{1n}^{n)} & b_1^{(n)} \\ 0 & a_{22}^{(n)} & \cdots & a_{2n}^{(n)} & b_2^{(n)} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right)$$

and the last step is to perform *backward substitution* in order to compute the solution:

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$

$$x_{n-1} = \frac{b_{n-1}^{(n)} - a_{n-1,n}^{(n)} x_n}{a_{nn}^{(n)}};$$

and generally,

$$x_j = \frac{b_j^{(n)} - \sum_{i=j+1}^{n} a_{j,i}^{(n)} x_i}{a_{jj}^{(n)}}$$

for $j = n-2, n-3, \cdots, 1$.

The diagonal elements $a_{kk}^{(k)}$ are called *pivot* elements. If $a_{kk}^{(k)} = 0$ for some k the multipliers are not defined. A remedy for such a complication is to exchange the kth row with another row below it, say row j for which $a_{jk}^{(k)} \neq 0$. This is an elementary row operation and the solution of the system doesn't change. The procedure of swapping rows to avoid a zero element on the diagonal is called *pivoting*. There are various kinds of pivoting, we will demonstrate here the *partial pivoting* which consists in the following:

If $a_{kk}^{(k)} = 0$ search the elements of the kth column below the diagonal element $a_{kk}^{(k)}$ and select the element, say on the ith row, for which $|a_{ik}^{(k)}|$ is the maximal one. Then interchange rows k and i and proceed with elimination.

The partial pivoting is based on the following theoretical result:

Theorem. Let A be nonsingular. If $a_{kk}^{(k)} = 0$, there is an $i > k$ for which $a_{ik}^{(k)} \neq 0$.

Example: Consider the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + x_2 + 2x_3 &= 2 \\ x_1 + 2x_2 + 2x_3 &= 2 \end{aligned}$$

The augmented matrix is

$$(A|b) = \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 \end{array}\right).$$

We perform the Gaussian elimination for the first column obtaining

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 \end{array}\right) \begin{array}{l} \\ m_{21} = 1 \\ m_{31} = 1 \end{array} \Longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right).$$

The pivot element $a_{22}^{(1)} = 0$ hence we need to interchange the second row with the (only) row under it which is the third row obtaining

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right).$$

The third row is already in the desired form, there is no need to iterate further. The solution is obtained by backward substitution:

$$x_3 = 1, \quad x_2 = 0, \quad x_1 = 0.$$

2. LU factorization

If we use the Gaussian elimination to find the solution of $Ax = b$, we can use the information gathered through the iterations to solve, by using just backward and forward substitutions, systems $Ax = b'$ with the same A but a different b'. Specifically, the Gaussian elimination makes the transformation:

$$A \Rightarrow U = \begin{pmatrix} a_{11}^{(n)} & a_{12}^{(n)} & \cdots & a_{1n}^{(n)} \\ 0 & a_{22}^{(n)} & & \\ & & \ddots & \vdots \\ 0 & \cdots & & a_{nn}^{(n)} \end{pmatrix}$$

Theorem. If the Gaussian elimination results in the upper triangular matrix U without pivoting, then the matrix A is equal to the product of the two matrices L and U,

$$A = LU,$$

where the matrix L is defined in the following way:

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & & \\ & & \ddots & \vdots \\ m_{n1} & \cdots & m_{n.n-1} & 1 \end{pmatrix}$$

Based on this factorization, we can solve the system $Ax = b$ by solving two triangular systems in the following way: solve first $Ly = b$ to obtain y and then $Ux = y$ to get the solution. That is,

$$Ux = y, \quad Ly = b \quad \Leftrightarrow \quad LUx = Ly = b \quad \Leftrightarrow \quad Ax = b.$$

Example: Consider

$$\left(\begin{array}{ccc|c} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & 1 \end{array}\right).$$

1) $m_{21}^{(1)} = a_{21}/a_{11} = -1/2, m_{32}^{(1)} = a_{31}/a_{11} = 0.$

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \begin{matrix} m_{21} = -1/2 \\ m_{31} = 0 \end{matrix} \quad \Longrightarrow \begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

2) $m_{32}^{(2)} = a_{32}^{(2)}/a_{22}^{(2)} = -2/3$

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad m_{32}^{(2)} = -2/3 \quad \Longrightarrow \begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{pmatrix} = U.$$

We have L in the form

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & -2/3 & 1 \end{pmatrix}$$

One can check directly that

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & -2/3 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{pmatrix} = LU.$$

To solve the initial equation, we first solve

$$Ly = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & -2/3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = b = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

by backward substitution, obtaining

$$y_1 = 1, \quad y_2 = 1/2, \quad y_3 = 4/3.$$

To get the solution x of the initial equation we solve

$$Ux = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 4/3 \end{pmatrix}$$

to obtain

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

What if we need pivoting?

We can still obtain a LU factorization, but of a transformation of the matrix A taking into account the interchanging of rows made through pivoting. We demonstrate this on the example we discussed before.

Example:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 3 \end{array}\right) \begin{array}{c} \\ m_{21}=1 \\ m_{31}=1 \end{array} \Longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{array}\right).$$

We need pivoting here since $a_{22}^{(1)} = 0$; we interchange the second row with the third row obtaining

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \end{array}\right).$$

The third row is already in the desired form, hence $m_{32}^{(2)} = 0$. The matrix L is then in the form

$$L = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right).$$

The product LU gives us

$$LU = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right) \left(\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right) = \left(\begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 2 \end{array}\right).$$

The resulting matrix is the one obtained from A by interchanging the second and the third rows: exactly what we did when pivoting.

We can still use the LU factorization for solving the initial equation if we do the same with b what we did with the matrix A: interchange the second and the third column; that is, the second and the third components. By solving $Lu = b$ by forward substitution we obtain

$$\left(\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right) \left(\begin{array}{c} y_1 \\ y_2 \\ y_3 \end{array}\right) = \left(\begin{array}{c} 1 \\ 3 \\ 2 \end{array}\right) \Rightarrow y = \left(\begin{array}{c} 1 \\ 2 \\ 1 \end{array}\right).$$

The equation $Ux = y$ gives us by backward substitution

$$\left(\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{array}\right) \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right) \Rightarrow x = \left(\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right).$$

There are several classes of matrices that never need pivoting. We will mention here only one such class: the matrices that are *strictly diagonally dominant*. A matrix A with components a_{ij} is strictly diagonally dominant when

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad \text{for all} \quad i = 1, 2, \cdots, n.$$

Such a matrix we will meet when we calculate the values of the second derivatives at the nodes of the natural cubic splines over uniform grid:

$$\left(\begin{array}{ccccccc} 4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 4 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 4 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & & \cdots & 1 & 4 \end{array}\right).$$

All the diagonal elements are equal to 4 and the sum of the other ones is 2. This means that when we compute cubic splines we do not need pivoting. Similar matrices arises also in finite-difference schemes for solving differential equations.

3. Error Analysis and Conditioning

Example: Consider

$$\left(\begin{array}{cc|c} \varepsilon & 1 & 1+\varepsilon \\ 1 & 1 & 2 \end{array}\right), \qquad (*)$$

where ε denotes a number much smaller than 1, say, $\varepsilon = 1. \times 10^{-6}$. This means that, rounding with 5 significant digits (any mantissa), we have $\varepsilon = 1. \times 10^{-6}, 1. + \varepsilon = 0.100001 \times 10^{1} \approx 0.1 \times 10^{1} = 1., 1 - 1/\varepsilon = 1 - 1,000,000 = -999,999 \approx= -0.1 \times 10^{6} = -1/\varepsilon$.

Let's apply Gaussian Elimination to the above system. The only multiplier we have to find is $m_{21} = 1/\varepsilon$. The resulting augmented matrix is

$$\left(\begin{array}{cc|c} \varepsilon & 1 & 1+\varepsilon \\ 0 & 1-1/\varepsilon & 1-1/\varepsilon \end{array}\right), \qquad (**)$$

The exact solution of the system $(*)$ above is $x_1 = 1, x_2 = 1$. If we take into account the roundoff error in $(**)$, we obtain the system

$$\left(\begin{array}{cc|c} \varepsilon & 1 & 1 \\ 0 & -1/\varepsilon & -1/\varepsilon \end{array}\right),$$

the solution of which is $x_1 = 0, x_2 = 1$.

In this example, the Gaussian elimination leads to an equation in which a small change, due to roundoff error, of the components of the matrix A and the vector b, leads a significant change in the solution. This of course is an academic example only but it illustrates that we have to be aware of the influence of errors of the data on the solution.

Intermission: Matrix Norms

Recall that a *vector norm* in the space of vectors $\mathbf{R}^n$ is a function, the value of which at x we denote $\|x\|$ with the following properties:

(1) $\|x\| \geq 0, \|x\| = 0 \iff x = 0$;

(2) $\|\alpha x\| = |\alpha| \|x\|$ for every scalar α;

(3) $\|x + y\| \leq \|x\| + \|y\|$.

Common examples of vector norms are the Euclidean 2-norm

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

and the so-called *infinity* norm

$$\|x\|_\infty = \max_i |x_i|.$$

For a given vector norm $\|\cdot\|$, one defines the *associated matrix norm* of a square matrix A in the following way:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

From this definition we obtain the following inequality:

$$\|Ax\| \leq \|A\|\|x\|.$$

Also, another consequence of the definition is the inequality

$$\|AB\| \leq \|A\|\|B\|$$

for any two matrices A and B. Indeed

$$\|AB\| = \max_{x \neq 0} \frac{\|ABx\|}{\|x\|} \leq \max_{x \neq 0} \frac{\|A\|\|Bx\|}{\|x\|} \leq \max_{x \neq 0} \frac{\|A\|\|B\|\|x\|}{\|x\|} = \|A\|\|B\|.$$

The matrix norm $\|A\|_2$ which is associated with the 2-norm of vectors is

$$\|A\|_2 = \sqrt{|\lambda_{\max}|},$$

where $\lambda_{\max}$ is the eigenvalue of $A^T A$ with maximal absolute value:

$$|\lambda_{\max}| = \max_i \{|\lambda_i| : \lambda_i \text{ is an eigenvalue of } A^T A\}.$$

The matrix infinity norm can be computed as follows:

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Example: Consider

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix}.$$

Then

$$A^T A = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}.$$

The eigenvalues are the roots of the equation

$$\det(\lambda I - A) = \det \begin{pmatrix} \lambda - 4 & -2 \\ -2 & \lambda - 2 \end{pmatrix} = \lambda^2 - 6\lambda + 4 = 0.$$

The roots are $3 \pm \sqrt{5}$ and hence $\|A\|_2 = \sqrt{3 + \sqrt{5}}$. The infinity norm is $\|A\|_\infty = \max\{0+1, 2+1\} = 3$.

4. Condition Number

Consider the equation

$$Ax = b$$

with a nonsingular matrix A and suppose that the vector b in right hand side changes to $b + r$. The perturbation r is often called *residual*. As a result, we will have a new solution which we denote $x + e$, where e is the *error* in the solution:

$$A(x + e) = b + r.$$

From the above two equations we obtain

$$Ae = r \quad \Rightarrow \quad e = A^{-1}r,$$

hence

$$\|e\| \leq \|A^{-1}\|\|r\|. \tag{1}$$

We also know that $b = Ax$, hence

$$\|b\| = \|Ax\| \leq \|A\|\|x\| \quad \Rightarrow \quad \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{2}$$

The product of the left sides of (1) and (2) will be than less or equal to the product of the right sides, that is,

$$\frac{\|e\|}{\|x\|} \leq \|A^{-1}\|\|r\|\frac{\|A\|}{\|b\|} = \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}.$$

The left hand side of this inequality if the *relative error*, and the inequality means that the relative error is bounded by a certain number depending on the matrix A only times the *relative residual*. This number is of great importance in numerical linear algebra.

Definition. The product $\|A\|\|A^{-1}\|$ is called *condition number* of the matrix A,

$$\mathrm{cond}(A) = \|A\|\|A^{-1}\|.$$

In the lines above the definition we derived the following fundamental result:

Theorem. The error of a solution x to the system $Ax = b$ is bounded by the condition number times the relative residual:

$$\frac{\|e\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|r\|}{\|b\|}.$$

Example.

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix}.$$

Take the ∞ norm. We have

$$\|A\|_\infty = \max\{1, 3\} = 3.$$

We have

$$A^{-1} = \begin{pmatrix} -1/2 & 1/2 \\ 1 & 0 \end{pmatrix}$$

hence

$$\|A^{-1}\|_\infty = \max\{1, 1\} = 1$$

and then

$$\text{cond}A_\infty = 3.$$

This means that, for example, if the relative residual is 10%, then the relative error will be not more than 30%.

Example.

$$A = \begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}.$$

We have

$$\|A\|_\infty = 2.$$

We have

$$A^{-1} = \begin{pmatrix} 25.25 & -24.75 \\ -24.75 & 25.25 \end{pmatrix}$$

hence

$$\|A^{-1}\|_\infty = 50$$

and then

$$\text{cond}A_\infty = 100.$$

In this case the relative residual (in the infinity norm) may be multiplied 100 times in the relative error (in the infinity norm); e.g., for relative residual 1% the relative error could be 100%.

Another fundamental result involving the condition number is the so-called *distance to singularity* formula which is due to Eckart and Young. It says the following:

Theorem.

$$\min\left\{\frac{\|B\|}{\|A\|} : A + B \text{ singular}\right\} = \frac{1}{\text{cond}(A)}.$$

Example. Can a 2% perturbation in the infinity norm of the matrix

$$A = \begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}.$$

make the matrix $A + B$ singular?

We already computed the condition number; $\text{cond}A_\infty = 100$. According to the theorem above, the distance to singularity is $1/100 = 0.01$. This means that a perturbation of A of the form $A + B$ with

$$.02\frac{\|B\|_\infty}{\|A\|_\infty} > 0.01$$

can make $A + B$ singular. Hence the answer to the question is "YES". Indeed, if we take the matrix

$$B = \begin{pmatrix} 0 & .02 \\ .02 & 0 \end{pmatrix}$$

with $\|B\|_\infty = .02$, and

$$\frac{\|B\|_\infty}{\|A\|_\infty} = \frac{.02}{2} = .01$$

we obtain

$$A + B = \begin{pmatrix} 1.01 & 1.01 \\ 1.01 & 1.01 \end{pmatrix}.$$

which is a singular matrix. Thus, even a 1% perturbation can make A singular.

Gaussian elimination and conditioning

Consider the example

$$\left(\begin{array}{cc|c} \varepsilon & 1 & 1+\varepsilon \\ 1 & 1 & 2 \end{array}\right). \tag{3}$$

Applying the Gaussian elimination we use the multiplier $m_{21} = 1/\varepsilon$ and arrive at the system

$$\left(\begin{array}{cc|c} \varepsilon & 1 & 1+\varepsilon \\ 0 & 1-1/\varepsilon & 1-1/\varepsilon \end{array}\right). \tag{4}$$

The ∞-norm condition number of the matrix in (3) is

$$\mathrm{cond}(A) \approx 4,$$

while the ∞-norm condition number of $A^{(1)}$ in (4) is

$$\mathrm{cond}(A^{(1)}) \approx \frac{1}{\varepsilon^2}$$

The condition numbers may grow with the iterations and then small perturbations (round-off errors) may lead to big errors. To avoid this, one uses pivoting even though the pivoting element is nonzero.

For the example in (3), by exchanging the rows we obtain

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ \varepsilon & 1 & 1+\varepsilon \end{array}\right)$$

which, with the multiplier $m_{21} = \varepsilon$ gives us

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1-\varepsilon & 1-\varepsilon \end{array}\right)$$

The condition number of the new matrix is close to 4, not much different from the first one. This observation is true in much more general situation: the partial pivoting stabilizes the condition number.

Intermission: Eigenvalues

The complex number λ is an eigenvalue of the square matrix A when λ is a solution of the equation

$$(*) \qquad \det(\lambda I - A) = 0.$$

The equation $(*)$ is called the *characteristic polynomial.*

The an eigenvalue is not reals, $\lambda = a + ib$, then the conjugate number $\lambda = a - ib$ is also an eigenvalue. If A is symmetric, all its eigenvalues are real numbers. A is singular if and only if the zero is an eigenvalue of A.

Examples.

1)

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

$$\det(\lambda I - A) = \det \begin{pmatrix} \lambda - 1 & -2 \\ -3 & \lambda - 4 \end{pmatrix} = (\lambda - 1)(\lambda - 4) - 6 = 0.$$

There are 2 real eigenvalues: $\lambda_1 = (5 + \sqrt{33})/2$ and $\lambda_1 = (5 - \sqrt{33})/2$.

1)

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

$$\det(\lambda I - A) = \det \begin{pmatrix} \lambda & -1 \\ 1 & \lambda \end{pmatrix} = \lambda^2 - 1 = 0.$$

There are 2 complex eigenvalues: $\lambda_1 = i$ and $\lambda_1 = -i$ that are conjugate.

5. Iterative Methods

The idea of the iterative methods for solving linear equations $Ax = b$ is the same as the fixed point iteration: first convert the equation to a fixed-point problem

$$x = Bx + c \tag{1}$$

with appropriate B and c and then apply the iteration procedure

$$x^{(k+1)} = Bx^{(k)} + c \tag{2}$$

with a given $x^{(0)}$.

The main problem in such a procedure is the question of convergence; whether the sequence $x^{(0)}, x^{(1)}, x^{(2)}, \cdots, x^{(n)}, \cdots$ converges to a solution x^*. Answer to this question is given by the theoretical results below in which one uses the concept of spectral radius.

Definition. The set of all eigenvalues of a matrix A is called *spectrum* and is denoted $\sigma(A)$. The number

$$\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$$

is called *spectral radius* of the matrix A.

Example: The matrix

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

has 2 complex eigenvalues: $\lambda_{1,2} = \pm i$ and $\rho(A) = 1$.

For the spectral radius of a matrix A the following inequality holds:

$$\rho(A) \leq \|A\| \quad \text{for any norm .}$$

Theorem. If $\rho(B) < 1$ then the fixed-point problem (1) has a unique solution x^* and for every initial point $x^{(0)}$ the sequence $x^{(n)}$ generated by the method (2) is convergence to x^*.

Corollary. If there exists a norm $\|\cdot\|$ for which $\|B\| < 1$, then the conclusions of the theorem above hold.

The different iteration methods differ in how the initial equation $Ax = b$ is converted to a fixed point problem $x = Bx + c$.

Richardson Iteration

The idea of the Richardson method is to convert the linear equation $Ax = b$ into the following fixed point problem:

$$x = (I - A)x + b \tag{1}$$

and then apply the iteration procedure

$$x^{(k+1)} = (I - A)x^{(k)} + b \tag{2}$$

with a arbitrarily chosen $x^{(0)}$. Of course, the convergence of this method will depend on the spectral radius of the matrix $(I - A)$.

Example: For the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1.5 \end{pmatrix}.$$

we have

$$I - A = \begin{pmatrix} 0 & -1 \\ 0 & 0.5 \end{pmatrix}.$$

which has eigenvalues: $\lambda_1 = 0$ and $\lambda_2 = .5$. Then $\rho(I - A) = .5$ and the Richardson iteration converges. Let's take $b = (1\ 0)$. Then the Richardson iteration has the form

$$\begin{aligned} x_1^{(k+1)} &= -x_2^{(k)} + 1 \\ x_2^{(k+1)} &= .5x_2^{(k)}. \end{aligned}$$

In order to improve the convergence, *parameterized* or *relaxed* Richardson iteration is used. The idea is to use the fixed point problem with a parameter,

$$x = (I - \alpha A)x + \alpha b$$

which is of course equivalent to $Ax = b$ for any $\alpha \neq 0$, and then to apply the fixed point iteration

$$x^{(k+1)} = (I - \alpha A)x^k + \alpha b$$

choosing α is such a way that the spectral radius of $I - \alpha A$ is as small as possible.

Example: For the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

we have

$$I - \alpha A = \begin{pmatrix} 2 - \alpha & -\alpha \\ 0 & 1 - \alpha \end{pmatrix}.$$

which has eigenvalues $\lambda_1 = 1 - \alpha, \lambda_2 = 2 - \alpha$. The spectral radius as a fucntion of α has the form

$$\rho(\alpha) = \min\{|1 - \alpha|, |2 - \alpha|\}$$

and $\min_\alpha \rho(\alpha) = 1/3$, attained for $\alpha = 1/3$.

For $b = (1\ 0)$ the relaxed Richardson iteration with $\alpha = 1/3$ will be has the form

$$\begin{aligned} x_1^{(k+1)} &= \frac{5}{3}x_1^{(k)} - \frac{1}{3}x_2^{(k)} + 1 \\ x_2^{(k+1)} &= \frac{2}{3}x_2^{(k)}. \end{aligned}$$

Jacobi method

In the Jacobi method the matrix A is split into two matrices,

$$A = D + C,$$

where D is the diagonal part of A and C is the rest. That is,

$$D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

and

$$C = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & \cdots & 0 \end{pmatrix}.$$

The initial equation is transformed into a fixed point problem in the following way:

$$Ax = b \quad \Rightarrow \quad (D + C)x = b \quad \Rightarrow \quad Dx = -Cx + b.$$

In terms of the fixed point problem (1) we have

$$B = -D^{-1}C, \quad c = D^{-1}b.$$

The Jacobi method has the following iteration:

$$Dx^{(k+1)} = -Cx^{(k)} + b.$$

The idea is at each iteration to solve a diagonal system; that is, n independent scalar equations. Of course, this system has to have a unique solution, that is, $\det(D) \neq 0$.

According to the general theory, the Jacobi method is convergent if

$$\rho(-D^{-1}C) < 1.$$

Example. Consider

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_1 + 2x_2 &= 1 \end{aligned}$$

with solution

$$x^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The Jacobi iteration has the form

$$\begin{aligned} 2x_1^{(k+1)} &= x_2^{(k)} + 1 \\ 2x_2^{(k+1)} &= x_1^{(k)} + 1. \end{aligned}$$

Starting from

$$x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

we have

$$\begin{aligned} 2x_1^{(1)} &= x_2^{(0)} + 1 \\ 2x_2^{(1)} &= x_1^{(0)} + 1 \end{aligned}$$

which results in

$$x^{(1)} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

The second and the third iterations give us

$$x^{(2)} = \begin{pmatrix} 3/4 \\ 3/4 \end{pmatrix} \quad x^{(3)} = \begin{pmatrix} 7/8 \\ 7/8 \end{pmatrix}.$$

Will the sequence converge? To answer this question we find the spectral radius of the matrix $-D^{-1}C$ with

$$D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

that is,

$$-D^{-1}C = -\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}.$$

The characteristic polynomial of the latter matrix is

$$\lambda^2 - \frac{1}{4} = 0.$$

The matrix $-D^{-1}C$ has 2 eigenvalues, both equal to $1/2$. The spectral radius is < 1, hence we have convergence.

Gauss-Seidel method

The Gauss-Seidel method uses the splitting of the matrix A into two matrices,

$$A = G + S,$$

where G is the lower triangular part of A including the main diagonal and S is the rest:

$$G = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

and

$$S = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \cdots & a_{2n} \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The transformation into a fixed point problem is carried out in the following way:

$$Ax = b \quad \Rightarrow \quad (G+S)x = b \quad \Rightarrow \quad Gx = -Sx + b.$$

In terms of the fixed point problem (1) we have

$$B = -G^{-1}S, \quad c = G^{-1}b.$$

The Gauss-Seidel iteration is of the form

$$Gx^{(k+1)} = -Sx^{(k)} + b.$$

The idea is at each iteration to solve a system of equations whose matrix is in a lower triangular form, that is, by forward substitution. The Gauss-Seidel iteration is convergent if

$$\rho(-G^{-1}S) < 1.$$

Theorem. If A is a positive definite matrix, then $\rho(-G^{-1}S) < 1$, and hence the GS method converges.

Example. Consider the same example as for the Jacobi method

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_1 + 2x_2 &= 1. \end{aligned}$$

The Gauss-Seidel iteration has the form

$$\begin{aligned} 2x_1^{(k+1)} &= x_2^{(k)} + 1 \\ -x_1^{(k+1)} + 2x_2^{(k+1)} &= 1. \end{aligned}$$

Starting from

$$x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

we have

$$\begin{aligned} 2x_1^{(1)} &= x_2^{(0)} + 1 \\ -x_1^{(1)} + 2x_2^{(1)} &= 1 \end{aligned}$$

which gives us

$$x^{(1)} = \begin{pmatrix} 1/2 \\ 3/4 \end{pmatrix}.$$

The point obtained from the second and the third iterations are

$$x^{(2)} = \begin{pmatrix} 7/8 \\ 15/16 \end{pmatrix} \quad x^{(3)} = \begin{pmatrix} 31/32 \\ 63/64 \end{pmatrix}.$$

From the form of the matrices G and S,

$$G = \begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix} \quad \text{and } S = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$$

we obtain that the spectral radius of $-G^{-1}S$ is $1/4 < 1$. Thus we have convergence.

In general, *the smaller the spectral radius, the faster the convergence.*